(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平7－36897

(43)公開日　平成７年(1995)２月７日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/27					
17/30					
		7315－5Ｌ	Ｇ０６Ｆ　15/ 20	５５０　Ａ	
		9194－5Ｌ	15/ 401	３１０　Ｄ	

審査請求　未請求　請求項の数３　ＯＬ　（全５頁）

(21)出願番号　特願平5－181264

(22)出願日　平成５年(1993)７月22日

(71)出願人　000005049
シャープ株式会社
大阪府大阪市阿倍野区長池町22番22号

(72)発明者　上田　徹
大阪府大阪市阿倍野区長池町22番22号　シャープ株式会社内

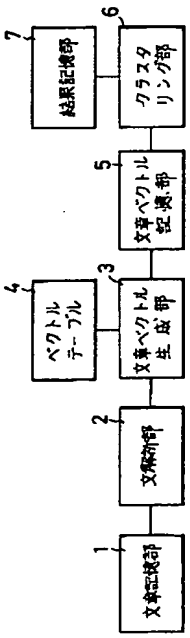
(74)代理人　弁理士　川口　義雄　（外１名）

(54)【発明の名称】　文書分類装置

(57)【要約】

【目的】　文書の自動分類を容易に行い得る文書分類装置を提供することにある。

【構成】　入力される文書に関して形態素解析を行う解析手段（１，２）と、解析された形態素のうち少なくとも自立語の特徴を示すベクトルを備えたベクトルテーブル（４）と、ベクトルテーブル（４）から自立語に対応するベクトルを抽出し抽出されたベクトルに基づいて文書の特徴を示す文章ベクトルを生成する文章ベクトル生成手段（３，５）と、生成された文章ベクトルを群分けし群分けされた文章ベクトルに基づいて文書を自動的に分類するクラスタリング手段（６，７）とを含むことを特徴とする。



1

【特許請求の範囲】

【請求項１】　入力される文書に関して形態素解析を行う解析手段と、解析された形態素のうち少なくとも自立語の特徴を示すベクトルを備えたベクトルテーブルと、ベクトルテーブルから自立語に対応するベクトルを抽出し抽出されたベクトルに基づいて文書の特徴を示す文章ベクトルを生成する文章ベクトル生成手段と、生成された文章ベクトルを群分けし群分けされた文章ベクトルに基づいて文書を自動的に分類するクラスタリング手段とを含むことを特徴とする文書分類装置。

【請求項２】　前記クラスタリング手段は群分けされた文章ベクトルから群を代表する代表ベクトルを算出し算出された代表ベクトルに基づいて文書を分類することを特徴とする請求項１に記載の文書分類装置。

【請求項３】　前記クラスタリング手段は代表ベクトルを構成する要素を抽出し抽出された要素に基づいて文書を分類することを特徴とする請求項２に記載の文書分類装置。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は文書分類装置に係り、詳細には文書を保存／自動分類する文書自動分類機やワープロ／ファイリングシステム等の分野に利用される文書分類装置に係る。

【０００２】

【従来の技術】従来、文書の自動分類は困難でありユーザが手動で分類を行ったり、文書中のキーワードを抽出し、あらかじめ作成されたシソーラスを用いて分類を行っていた。

【０００３】

【発明が解決しようとする課題】キーワードを抽出し、得られたキーワードの関係をシソーラスを用いて推定する場合、得られるのはキーワードと、他のキーワードとの関係であり文書と他の文書との関係ではない。よって、この方式での分類では分類精度が極めて悪い。本発明の目的は、文書の自動分類を容易に行い得る文書分類装置を提供することにある。

【０００４】

【課題を解決するための手段】入力される文書に関して形態素解析を行う解析手段と、解析された形態素のうち少なくとも自立語の特徴を示すベクトルを備えたベクトルテーブルと、ベクトルテーブルから自立語に対応するベクトルを抽出し抽出されたベクトルに基づいて文書の特徴を示す文章ベクトルを生成する文章ベクトル生成手段と、生成された文章ベクトルを群分けし群分けされた文章ベクトルに基づいて文書を自動的に分類するクラスタリング手段とを含むことを特徴とする。

【０００５】

【作用】解析手段が入力される文書に関して形態素解析を行い、ベクトルテーブルが解析された形態素のうち少

2

なくとも自立語の特徴を示すベクトルを備えており、文章ベクトル生成手段がベクトルテーブルから自立語に対応するベクトルを抽出し抽出されたベクトルに基づいて文書の特徴を示す文章ベクトルを生成し、クラスタリング手段が生成された文章ベクトルを群分けし群分けされた文章ベクトルに基づいて文書を自動的に分類するので、文章中に含まれる自立語からその文章の大体の意味を現す文章ベクトルが抽出され、その文章ベクトルを特徴としてクラスタリングが行われ、シソーラスを使用することなく複数の文章を内容に応じて自動的に分類し得る。

【０００６】

【実施例】図１は本発明の文書分類装置の実施例のブロック図、図２は本発明の文書分類装置の他の実施例のブロック図、図３は本発明の文書分類装置の別の実施例のブロック図である。

【０００７】図１において、１はあらかじめ入力された複数の文書を記憶する文章記憶部、２は文章記憶部１とともに解析手段を構成し、文章記憶部１に記憶されている複数の文書をひとつひとつ形態素解析を行う文解析部、３は入力された文章の形態素解析を行い得られた自立語の特徴を示すベクトルを集計して入力文章の特徴ベクトルとする文章ベクトル生成部であって、換言すれば、文解析部２で得られた自立語について後述するベクトルテーブルを参照し、あらかじめ作成されたベクトル表現された自立語（もしくは自立語および付属語）の特徴を示すベクトルを登録しているベクトルテーブル４にその要素が存在するならば、該当するベクトルと現文章ベクトルとの演算（例えば加算）を行う文章ベクトル生成部、５は文章ベクトル成生部３とともに文書ベクトル生成手段を構成し、文書毎に文章ベクトル生成部３で求められた文章ベクトルを記憶する文章ベクトル記憶部、６は文章ベクトル記憶部５で記憶されている文章ベクトルを用いてクラスタリングを行って入力文書に含まれる自立語（もしくは自立語および付属語）の特徴から複数の文書を自動的に分類するクラスタリング部、７はクラスタリング部とともにクラスタリング手段を構成し、クラスタリング部６によって文書がいくつかの群に分けられたので、その結果を記憶する結果記憶部である。なお、クラスタリングにはＫ－ｍｅａｎ法など種々の方法が存在するが、ここではその手法は問わない。

【０００８】図２の構成の記号は図１と同一のものは同じ番号を付してある。８は結果記憶部７に記憶されている分類された群毎に、その群のもつ特徴を算出する代表ベクトル算出部でありこれにより分類された群がもつ特徴をベクトルの形で表現できる。９は新たな文書の文章ベクトルと代表ベクトル算出部８で求められた代表ベクトルとの距離を求める距離計算部である。図２の実施例においては、クラスタリング部６、結果記憶部７、代表ベクトル算出部８および距離計算部９がクラスタリング

3

手段を構成する。新たな文書の文章ベクトルが求められると、そのベクトルと各分類群の代表ベクトルとの距離が求められ、新たな文書はどの分類に近いか判定されて最も近い距離の代表ベクトルをもつ群に分類される。

【０００９】図３の構成の記号は図１及び図２と同一のものは同じ番号を付してある。１０は各群の代表ベクトルから大きな値の要素を抽出する要素抽出部、１１はベクトルの各要素が持つ意味付けを示す要素辞書である。図３の実施例においては、クラスタリング部６、結果記憶部７、代表ベクトル算出部８、要素抽出部１０および要素辞書１１がクラスタリング手段を構成する。代表ベクトルの要素が大きいところを取り出して、その言語的な意味を要素辞書から抽出することで、各分類群がもつ特徴を言語的に表すことができる。

【００１０】以下本発明の実施例の作動を説明する。

【００１１】図４はあらかじめ作成されたベクトルテーブルの例を示す図、図５は「国会への証人喚問」という文章が含まれる場合の文章ベクトルの計算例を説明する図、図６は「国会の解散に伴う総選挙について」という文章が含まれる場合の文章ベクトルの計算例を説明する図、図７は「国際経済における貿易収支の影響」という文章が含まれる場合の文章ベクトルの計算例を説明する図、図８は「円高の及ぼす影響」という文章が含まれる場合の文章ベクトルの計算例を説明する図である。

【００１２】図４には「選挙」「経済」「国会」「貿易」「円高」の５単語の特徴ベクトル（５次元）が記載されている。図５のように「国会への証人喚問」という文章が入力されると、文解析部２において形態素解析が行われ、その文章に含まれる単語の中で図４のベクトルテーブルに登録されている単語（自立語）のベクトルが抽出される。抽出された単語からベクトルテーブルを用いて単語のベクトルが選ばれる。複数の単語が１文章から抽出された場合には単語のベクトルを平均化することで文章ベクトルが計算される（ただし、小数点は切り捨て）。図４の「国会」のベクトルは１２８１７であるから図５の計算結果は１２８１７、図４の「国会」のベクトルは１２８１７、「選挙」は１０５２８であるから図６の計算結果は両者の平均値１１６１７、すなわち（１＋１）÷２＝１、（２＋０）÷２＝１、（８＋５）÷２＝６、（１＋２）÷２＝１、（７＋８）÷２＝７（ただし、小数点以下切り捨て、以下同様の計算を行う）図４の「経済」は７７０２５、「貿易」は８９１０５であるから図７の計算結果は両者の平均値７８０１５、図４の「円高」は９５１０２であるから図８の計算結果は９５１０２である。このようにして各文章の１つの文章ベクトルが求められる。

【００１３】文章ベクトルが求められたならば、従来のクラスタリングの手法を適用する。図５から図８において、図５、図６の２つの文章ベクトルと図７、図８の２つの文章ベクトルとはそれぞれ計算結果が近いので４つ

4

の入力文章、「国会への証人喚問」「国会の解散に伴う総選挙について」「国際経済における貿易収支の影響」「円高の及ぼす影響」は２つの群、すなわち「国会への証人喚問」「国会の解散に伴う総選挙について」の群と、「国際経済における貿易収支の影響」「円高の及ぼす影響」の群とに分類される。

【００１４】図９は群「国会への証人喚問」「国会の解散に伴う総選挙について」の文章ベクトルの計算結果から代表ベクトルを計算する例を説明する図、図１０は群「国際経済における貿易収支の影響」「円高の及ぼす影響」の文章ベクトルの計算結果から代表ベクトルを計算する例を説明する図である。

【００１５】前記群の分類ができたならば、各分類群毎に代表ベクトルを求める。単純には各分類文に属する文章の文章ベクトルの平均をとることで代表ベクトルが計算できる。図９においては、「国会への証人喚問」の１２８１７と「国会の解散に伴う総選挙について」の１１６１７との平均１１７１７が代表ベクトルとして求められる（小数点は切り捨て、以下同様）。また図１０においては、「国際経済における貿易収支の影響」の７８０１５と「円高の及ぼす影響」の９５１０２との平均８６００３が代表ベクトルとして求められる。

【００１６】あらたに未知の文章が入力された場合には、その文章ベクトルを求め、文章ベクトルと各文章ベクトルとの距離を計算することで、未知の文章と各分類群との距離が求められる。未知の文章を最も距離の近い分類群に入れることにより、以前の分類結果を保存したままあらたな文章を最も近い分類群に入れることができる。

【００１７】ベクトルの各要素にその要素がもつ意味を要素辞書として記憶しておくと、代表ベクトルから分類群のもつ特徴が得られる。図９の例で第５要素が政治に関連したものであるとしたなら、図９の例では第５要素が大きいため、この分類群は政治に関連した群であるということが分かる。

【００１８】

【発明の効果】解析手段が入力される文書に関して形態素解析を行い、ベクトルテーブルが解析された形態素のうち少なくとも自立語の特徴を示すベクトルを備えており、文章ベクトル生成手段がベクトルテーブルから自立語に対応するベクトルを抽出し抽出されたベクトルに基づいて文書の特徴を示す文章ベクトルを生成し、クラスタリング手段が生成された文章ベクトルを群分けし群分けされた文章ベクトルに基づいて文書を自動的に分類するので、文章中に含まれる自立語からその文章の大体の意味を現す文章ベクトルが抽出され、その文章ベクトルを特徴としてクラスタリングが行われ、シソーラスを使用することなく複数の文章を内容に応じて自動的に分類し得る。

【図面の簡単な説明】

5

【図１】本発明の文書分類装置の実施例のブロック図である。

【図２】本発明の文書分類装置の他の実施例のブロック図である。

【図３】本発明の文書分類装置の別の実施例のブロック図である。

【図４】あらかじめ作成されたベクトルテーブルの例を示す図である。

【図５】国会への証人喚問という文章が含まれている場合の文章ベクトルの計算例を説明する図である。

【図６】国会の解散に伴う総選挙についてという文章が含まれている場合の文章ベクトルの計算例を説明する図である。

【図７】国際経済における貿易収支の影響という文章が含まれている場合の文章ベクトルの計算例を説明する図

6

である。

【図８】円高の及ぼす影響という文章が含まれている場合の文章ベクトルの計算例を説明する図である。

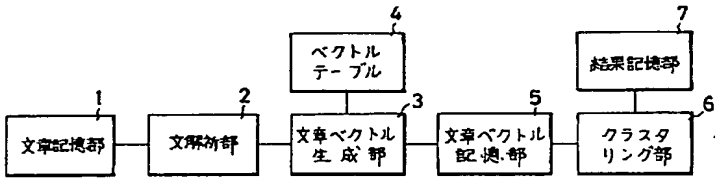
【図９】群の文章ベクトルの計算結果から代表ベクトルを計算する例を説明する図である。

【図１０】他の群の文章ベクトルの計算結果から代表ベクトルを計算する例を説明する図である。

【符号の説明】

1　文章記憶部
2　分解析部
3　文章ベクトル生成部
4　ベクトルテーブル
5　文章ベクトル記憶部
6　クラスタリング部
7　結果記憶部

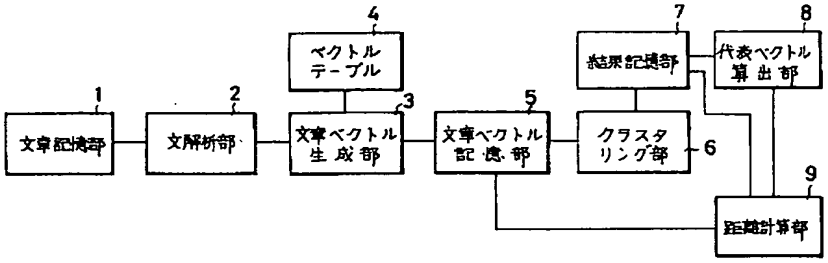
【図１】



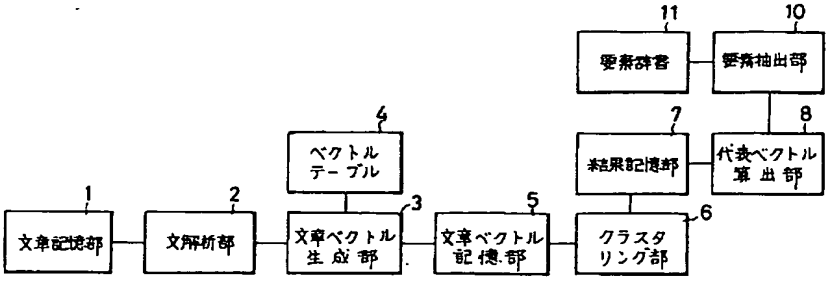
【図４】

選挙	1	0	5	2	8
経済	7	7	0	2	5
国会	1	2	8	1	7
貿易	8	9	1	0	5
円高	9	5	1	0	2

【図２】

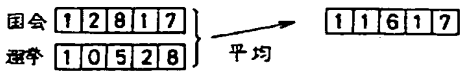


【図３】

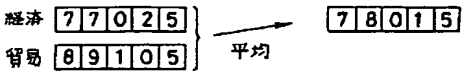


【図５】

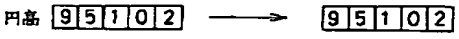
【図６】



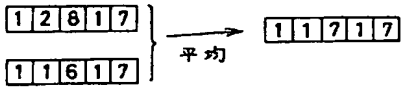
【図７】



【図８】



【図９】



【図１０】